7-1-2009

Prediction of Disorder with New Computational Tool: BVDEA

Irem Ersoz Kaya
Mersin University

Turgay Ibrikci
Cukurova University

Okan K. Ersoy
Purdue University - Main Campus, ersoy@purdue.edu

Prediction of Disorder with New Computational Tool: BVDEA

Irem Ersöz Kaya[1], Turgay Ibrikçi[2] and Okan K. Ersoy[3]
[1]Department of Electronics and Computer, Tarsus Technical Education Faculty,
Mersin University, Tarsus, Mersin, Turkiye
[2]Department of Electrical and Electronics Engineering, Çukurova University, Balcalı, Adana, Turkiye
[3]School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907

ABSTRACT

Motivation: Recognizing that many intrinsically disordered regions in proteins play key roles in vital functions and also in some diseases, identification of the disordered regions has became a demanding process for structure prediction and functional characterization of proteins. Therefore, many studies have been motivated on accurate prediction of disorder. Mostly, machine learning techniques have been used for dealing with the prediction problem of disorder due to the capability of extracting the complex relationships and correlations hidden in large data sets.

Results: In this study, a novel method, named Border Vector Detection and Extended Adaptation (BVDEA) was developed for predicting disorder as an alternative accurate classifier. The classifier performs the predictions by using three types of structural features belonging to proteins. For attesting the performance of the method, three computational learning techniques and eleven specific tools were used for comparison. Training was executed based on the data by 5-fold cross validation. When compared with the three learning methods of GRNN, LVQ and BVDA, the proposed method gives the best accuracy on classification. The BVDEA also provides faster and more robust learning as compared to the others. The new method provides a significant contribution to predicting disorder and order regions of proteins.

1 INTRODUCTION

The aim of this study is to find an efficient computational method that can provide information about the structural class among ordered and disordered proteins concerning different biochemical and physical features of amino acids. For this purpose, a new algorithm was proposed in the study, named Border Vector Detection and Extended Adaptation (BVDEA) to achieve an accurate prediction of disordered data. In this section, an overview of the studies on ordered and disordered proteins is given. The data sets and methods used are described in Section 2. Results and discussions are highlighted in Section 3, and Section 4 covers conclusions.

Widely believed dogma of structural biology states that the three dimensional (3D) structure of a protein is a prerequisite for its biological function. The tenet has been arisen more than 100 years ago with Fischer's "lock and key" model or Koshland's "induced fit" theory in which both the enzyme and its substrate were fit to each other with the complementary 3D shapes like a lock and key in order to fulfill enzymatic behaviour (Fischer, 1894; Koshland, 1958). The functional form of a protein is named as "the native state"; on the other hand, the loss of functional activity exhibits "the denatured state" of the protein. Generally the loss of function is associated with the lack of specific 3D structure (Mirsky and Pauling, 1936)

At the end of 1970s, research on structural/functional analysis indicated that some protein segments remain unfolded in their native states. Contrary to the structure – function paradigm, several protein regions keep failing to fold into a fixed 3D structure under physiological condition, yet exhibit function as discovered by experimental methods. Over the years, many proteins have been found to contain these unstructured regions and to play crucial role in a variety of functions including DNA binding, cell signaling and protein modification (Wright and Dyson, 1999). It has been shown that proteins can also be entirely disordered. Currently these proteins are generally referred to as "natively unfolded" or "intrinsically disordered" (Dunker *et al.*, 2000).

In the prior work of Dunker and colleagues, they pointed out the functional importance of the types of proteins (Dunker *et al.*, 1997). Moreover, many studies have also emphasized the importance and the necessity of the intrinsically disordered proteins for functionality. For instance, it has been elucidated that their flexible conformation enables them to form complexes with several different targets via disorder-to-order transitions upon binding as in enzyme binding with substrate, and protein binding with RNA/DNA/protein (Romero *et al.*, 1997; Obradovic *et al.*, 2003).

Recently, it was confirmed that some disordered proteins can also be involved in several diseases such as Alzheimer disease, Parkinson disease and certain types of cancer (Williams *et al.*, 2001; Galzitskaya *et al.*, 2006) Therefore, studies on structural identification of intrinsically disordered proteins have been thought to be an aid in drug design, protein expression and functional recognition. Due to their significance, dealing with structural identification of the proteins that was named as "The Protein Non-Folding Problem" has gained growing interest in structural bioinformatics (Li *et al.*, 2000).

In accordance with the studies on protein folding problem, disordered regions have also been identified by using experimental methods. In X-ray diffraction, missing electron density indicates disorder. In NMR, disorder is revealed by sharp peaks or the negative values of N-H heteronuclear Nuclear Overhauser Effect (NOE) measurements. CD uses UV spectra in order to identify disorder by low intensity wavelength. The method does not give residue-by-residue basis information but provides secondary structure estimation (Tompa *et al.*, 2002). Each method has its own pros and cons in characterization of disorder due to the quality of being able to examine different aspects of protein structure. For instance, because of its lack of regular secondary structure, a loopy protein can be identified as a disordered protein by CD spectroscopy analysis. A wobbly ordered region can be misclassified in X-ray crystallography. NMR has difficulties in discerning the molten globules from the random coils (Romero *et al.*, 2001).

Because of the constraints, it is desirable to confirm or to reinforce the experimental results by multiple methods. Besides, experimental methods are also quite expensive and time-consuming. Thus, alternative to experimental methods, several computational methods have been suggested for disorder prediction based on the amino acid sequence. Mostly preferred machine learning techniques can be cited as neural networks (Li *et al.*, 1999; Linding *et al.*, 2003a; Vucetic *et al.*, 2003; Radivojac *et al.*, 2004; Cheng *et al.*, 2005; Peng *et al.*, 2005) and support vector machines (Shimizu *et al.*, 2004; Weathers *et al.*, 2004; Peng *et al.*, 2006). In the majority of these studies, the input patterns have been mostly derived from a variety of sequence properties such as flexibility, amino acid frequency, complexity, charge, and secondary structure that characterize disorder.

To improve the quality of prediction, it has been recently made efforts to find more useful features and to develop more robust predictors (Garbuzynskiy *et al.*, 2004; Dosztányi *et al.*, 2005; Shimizu *et al.*, 2007). For instance, Jones et al. demonstrated in their study that using the position specific scoring matrices (PSSMs) within a defined length of window can improve the accuracy of predicting its disorder attribute (Jones and Ward, 2003).

Eventually, specific disorder prediction tools have been developed such as PONDRs, DisEMBL, GlobPlot, DISOPRED2, FoldIndex, RONN, DisPRO, PreLink, DisPSSMP.

The first tool designed specifically for prediction of protein disorder was PONDR (Predictor of Naturally Disordered Regions) called XL1 (Romero *et al.*, 1997). The trained feedforward neural network model achieved 58% prediction accuracy by using only 7 X-ray characterized partially disordered proteins. Prediction was based solely on the frequencies of 8 amino acid sequences (His, Glu, Lys, Ser, Asp, Cys, Trp and Tyr) and two average attributes. Subsequently, a series of PONDRs were constructed, and overall prediction accuracy ultimately exceeded 80%. PONDR VLXT reached 70% prediction accuracy by integrating their three initial neural network predictors, VL1, XN and XC. VL1 was trained on 25 variously characterized disordered proteins and gave an accuracy of 64% (Romero *et al.*, 2001). XN and XC were proposed as N- and C- terminal predictors (Li *et al.*, 1999). The overall accuracy of the most recent PONDR predictors (VL3 series) ranges from 79% to 83% with a Win size of 41 and Wout sizes ranging from 1 to 121 (Peng *et al.*, 2005). They attained this improvement by adding the feature of evolutionary knowledge.

DisEMBL provides a method based on artificial neural networks trained for predicting three different definitions of disorder. The method was used with all three types of disorder which are hot loops, coils and REMARK465, named as EMBL(hot), EMBL(coil) and EMBL(465) (Linding *et al.*, 2003). REMARK465 contains the regions that lack electron density in crystal structure. Hot loops indicate highly mobile loops where coils are the regions that have no regular secondary structures. The most accurate result for the prediction of hot loops was correct prediction rate of 64%. The method is publicly accessible as a web service at http://dis.embl.de.

Linding et al. designed a model, called GlobPlot based on amino-acid propensities for disorder/globularity (Linding *et al.*, 2003b). They proposed a new scale for propensities, named Russell/Linding to be either in regular secondary structures (a-helices or b-strands) or outside of them ('random coil', loops, turns etc.). GlobPlot is a web service at http://globplot.embl.de that allows the user to plot the disorder/globularity tendency of a query protein. At a specificity of 88%, they obtained a sensitivity of 28%.

Linear support vector machines based on a local amino acid sequence was trained in the DISOPRED2 method (Ward *et al.*, 2004a). Missing residues from the electron density map were defined as disordered regions to form the data set. A sequence profile derived from a PSI-BLAST search was used to construct the input vector for each residue within the window length of 15. Testing results were given with a

~93% accuracy. DISOPRED2 can be downloaded from the web site at http://bioinf.cs.ucl.ac.uk/disopred/ (Ward *et al.*, 2004b).

Uversky and coworkers proposed that a low mean hydrophobicity combined with relatively high mean net charge is a significant indicator of disorder (Uversky *et al.*, 2000). The web-based implementation of the algorithm, named as FoldIndex is served at http://bip.weizmann.ac.il/fldbin/findex (Prilusky *et al.*, 2005). In the per-sequence prediction method, 30 proteins from among the 39 intrinsically disordered proteins were correctly assigned with the label of "disordered" while 15 thru 151 ordered proteins were correctly classified as "ordered". The overall accuracy was reported as 83%.

Yang et al. proposed a novel model, Regional Order Neural Network (RONN), for prediction of disordered region based on the alignment scores of the windowed sequences against a series of sequences of known folding state (Yang *et al.*, 2005). In RONN, the determined distances from a subset of well-characterized prototype sequences are used to train a neural network for classifying the query sequence as ordered or disordered. They created a data set through the proteins from BDP by applying some filtering applications. 80 proteins of the data set were reserved for testing; the remaining sequences were used for training and validation. Another test set was derived from 79 completely disordered proteins and 80 completely ordered proteins. The method achieved the testing accuracies of 84.9% and of 78.9% for the main testing set and for the other, respectively. RONN is available at http://www.strubi.ox.ac.uk/RONN.

DISpro is a 1D-recursive neural network method that involves the use of evolutionary information incorporating the predicted secondary structure and relative solvent accessibility (Cheng *et al.*, 2005). The proteins used for training and testing were extracted from the PDB. They used only the proteins that were solved by X-ray crystallography with a resolution better than 2.5 Å. All proteins were required to be at least 30 residues long. Disordered regions were determined according to missing residues of ATOM records in PDB. The sensitivity result was given as 75.4% and the specificity was 38.8%.

PreLink is a method that relies on compositional bias and low hydrophobic cluster content and is accessible at http://genomics.eu.org/spip/PreLink. In this method, a linker set was created by aligning the BDP proteins with its corresponding Swiss-Prot record and extracting non-aligned regions. The resulting segments were classified by their position as N-terminal, C-terminal or inner fragments. The prediction involved using three values obtained from the calculation of the amino acid distributions in structured and unstructured regions, to estimate the probability that a given sequence fragment be part of either a structured or an unstructured region, and of the distance to the nearest hydrophobic cluster of each ami-

no acid (Coeytaux and Poupon, 2005). PreLink correctly predicted 59.9% of the N-terminal fragments, 70.5% of the C-terminal fragments and 61.1% of the inner fragments.

Su et al. used condensed position specific scoring matrix profiles by merging associated columns of the matrix concerning the several physicochemical properties of amino acids, called PSSMP (Su *et al.*, 2006). The training data set that contained 336880 residues was collected from the Protein Data Bank and the DisProt Database. They achieved rather successful predictions for two distinct test sets via training a Radial Basis Function (RBF) neural network based on their proposed PSSMP patterns. In the study, the same testing sets were used as with the study of Yang et al. The sensitivity of 0.767 and the specificity of 0.848 were given for the prediction results of the main test set while the obtained scores on testing the other set were 0.825 and 0.765, respectively.

2 METHODS

2.1 Datasets

In this study, 3 protein sets from different studies were used for training, validation and blind testing processes.

The first data set, including 80 proteins, was organized by Yang et al. to test the prediction accuracy of Regional Order Neural Network / RONN (Yang *et al.*, 2005). This data set was generated through the analysis of all entries obtained by NMR and X-ray crystallography in the Molecular Structure Database/MSD. For each MSD entry, the residue of the sequence observed in ATOM coordinate records in PDB was aligned with both the SEQRES records providing protein sequences and corresponding UniProt sequences. In this way, the unseen regions could be detected. 7327 entries were obtained considering unobserved regions containing at least 5 consecutive residues. These entries were divided into two groups as the ones having more than 20 consecutive unobserved residues (long set, 1573) and the others (short set, 5754). The long set received 872 entries after the filtering processes like the extraction of multi component complexes, the preference of only the highest numbered sequence within the same 3-letter code entries, and the selection of the residues which were not observed in other chains for entries containing multiple chains. Then, by using CD-Hit program, the ones with more than 70% sequence identity among the remaining proteins were eliminated (Li *et al.*, 1999).

The second data set with 80 completely ordered proteins were obtained from PONDR® website by Yang et al. (retrieved in February 2003) (Yang *et al.*, 2005). Romero et al. developed this data set in

2001 for the PONDR studies from their non-redundant database, named as O_PDB_Select_25 that was created by using PDB_Select_25 database (Romero *et al.*, 2001). PDB_Select_25 is a database that contains one protein sequence representing each protein group in PDB (Hobohm and Sander, 1994). By extracting only the ordered sequences from this database, the O_PDB _Select_25 database was created.

For the third data set, 79 out of 91 disordered proteins reported in the study of Uversky et al. in 2000 were used in our research. Uversky et al. reported 91 completely disordered proteins defined through spectroscopic methods in the literature (Uversky *et al.*, 2000). These proteins ranging in size between 50 and 1827 were stated in the literature as having the chemical shifts of a random-coil in nuclear magnetic resonance and/or lacking of regular secondary structure detected by circular dichroism/CD or Fourier transform infrared spectroscopy/FTIS and/or showing close hydrodynamic dimensions to a polypeptide chain under physiological conditions. In their studies, Uversky et al. pointed out that the proteins were not in ordered structure via low mean hydrophobicity and relatively high net charge relation.

The three data sets were named as D80, CO80 and CD79, respectively. The compositions of the sets are given in Table 1. From the 3 chosen protein sets, 2 data sets with different characteristics were constructed. The first data set was created by using D80 protein set which contains disordered regions and thus having an unbalanced distribution. This set was used as blind test set for comparing the success of the proposed method with existing predictors. The second set was developed by balancing equal amounts of completely ordered (CO) and completely disordered (CD) protein data sets, named as COD159.

2.2 Data Presentations

In order to develop successful models in estimating the protein structure, it is beneficial to consider amino acid neighboring in knowledge representation (Dunker *et al*., 2001). For this reason, the information about an amino acid is considered with respect to the amino acids surrounding it. According to this, all amino acids' knowledge in a *k* residue length window is included to represent the central amino acid of the window. Through sliding the window, all the residues in the protein are scanned in order to extract information for each sequence position (Qian and Sejnowski, 1988). However, inclusion of each amino acid individually in the window causes a curse of dimensionality (Bishop, 1995). Besides, using 20 bits of binary representation for these amino acids enlarges this problem. In order to avoid this dimensionality problem, it has been preferred to use the real value representations of amino acids and the average information of all the amino acids within the window (Peng *et al*., 2005).

Table 1. The compositions of the three data sets.

Data Set	D80	CO80	CD79
Number of Chains	80	80	79
Number of Ordered Regions	151	80	0
Number of Disordered Regions	183	0	79
Number of Ordered Residues	29909	16568	0
Number of Disordered Residues	3649	0	14462

In the way, each feature is represented by only one attribute in a pattern. This also removes the dependency on the window size. For n attributes, an n-dimensional pattern is created for a given position of a protein and labeled with the class of the corresponding amino acid at that position. Here, the class label is ordered or disordered. Thus, the estimation of disordered regions is considered as a binary classification problem. The per-residue prediction values obtained by the computational learning methods take the value of "1" for ordered and "0" for disordered.

In this study, new patterns of amino acids in the data sets were constructed in 3 different manners as training inputs for machine learning methods.

The first step for obtaining network input is to create initial profiles for each amino acid of all proteins in the data set by means of the sliding window technique. The centre of a window, with a size of an odd number, is placed targeting an amino acid and starting from the first residue; this is repeated by sliding one residue to the right until the end of sequence, thus all residues of a sequence are treated. In this study, the size of the sliding window was determined as 21 by considering the references (Linding *et al.*, 2003a; Dosztányi *et al.*, 2005). N and C terminals of proteins can cause bias as they have tendency to be disordered (Li *et al.*, 1999), thus the regions of the protein remaining outside the window with a number of (w-1)/2 amino acids from the beginning and from the end were excluded.

Previous studies have shown that several physicochemical properties of amino acids can be used for distinguishing disorder. Therefore, in the next step, by using property, composition and evolution values of the amino acids in the window, input profiles were constructed for each residue within a protein sequence. Then, these profiles were combined to derive an input pattern. The pattern attributes are named as $X(i)$ where i is the attribute index.

Finally, for each attribute, the values of the input patterns were re-scaled between 0-1 with min-max

normalization technique, namely,

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A \quad (1)$$

where v' is the new scale value for an attribute. $new\_\min_A$ and $new\_\max_A$ are the expected minimum and maximum values, respectively, equal to 0 and 1, and $\min_A$ and $\max_A$ are the minimum and maximum values of an attribute in the data set.

In the study, 49 property scales of amino acids from AAIndex and from different references were used for constituting property-based profiles. The contents and the references of all properties and their scale values are given in Appendix A.

The calculation of X_{iprop} attributes for ith position residue of a sequence with length N within the window of length w is given by

$$X_{iprop} = \frac{1}{w}\sum_{j=w_s}^{w_f} prop_j \quad (2)$$

where $w_s = i - (w-1)/2$, $w_f = i + (w-1)/2$, and $prop_j$ is the scale value of a given property for the amino acid at position j.

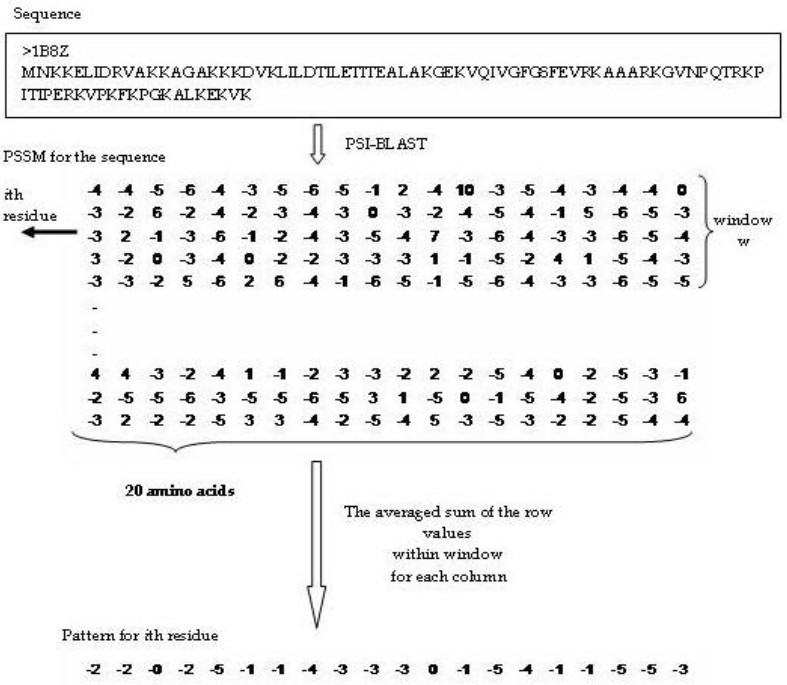


Fig. 1. The constructing of PSMM profile of a protein within a window.

As the final property-based attribute, the measure of sequence complexity called *K2* entropy was used (Wootton and Federhen, 1993). Complexity is a measure which shows how many different ways a sequence can be rearranged (Weathers *et al.*, 2004). It was demonstrated that low complexity regions are more likely to be disordered than ordered (Romero *et al.*, 2001; Peng *et al.*, 2005). Shannon's *K2* entropy, X_{iK2} value over a window was calculated from the 20 amino acid frequencies, X_{ia} within the window, as

$$X_{iK2} = -\sum_{a=1}^{20} X_{ia} \log_2 X_{ia} \tag{3}$$

Consequently, 50 attributes of an input pattern were derived from the property-based profile.

In the construction of composition-based profiles, first order statistics of 20 known amino acids and compositions for 30 different property groups of amino acids within the window were examined.

The presence of each known 20 amino acids within the window, $a \in$ {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, constitutes the first 20 attributes of the profile. The next 30 attribute values of the profile are derived from the frequency of the specified amino acids belonging to the same property groups in a given window (Appendix A). A total of 50 attributes constitutes a composition-based profile.

For the *i*th position residue of a sequence with length *N*, the calculation of X_{ia} attribute value corresponding to an amino acid within the window of length *w* is given by

$$X_{ia} = \frac{1}{w} \sum_{j=w_s}^{w_f} P_{ja} \tag{4}$$

where $w_s = i - (w-1)/2$ and $w_f = i + (w-1)/2$, $P_{ja} = 1$ if amino acid *a* is at position *j*, otherwise $P_{ja} = 0$.

Similarly, an attribute value of a group composition is calculated by averaging the numbers of amino acids which pertain to a given property group within the window. For instance, the sum of X_{ia} values of amino acids which have hydrophobicity feature, $a \in$ {A, I, L, M, F, P, W, V, G}, provides the X_i (property) attribute value of the property group (5):

$$X_i(hydrophobicity) = X_{iA} + X_{iI} + X_{iL} + X_{iM} + X_{iF} + X_{iP} + X_{iW} + X_{iV} + X_{iG} \tag{5}$$

Here, the net charge value for a residue should be calculated considering the sign of that amino acid. Based on this, the number of negative charged amino acids, $a \in$ {D, E}, is subtracted from the number of positive charged amino acids, $a \in$ {K, R} (6):

$$X_i(net\ charge) = X_{iK} + X_{iR} - X_{iD} - X_{iE} \tag{6}$$

It was shown that the use of evolutionary knowledge improves the accuracy of disorder prediction (Jones and Ward, 2003; Ward *et al.*, 2004b; Su *et al.*, 2006). Therefore, position-specific scoring matrices (PSSM) generated by PSI-BLAST search were used to construct evolution-based profiles in the study. PSI-BLAST provides a measure of residue conservation in a given position (Altschul *et al.*, 1997) and returns a 20-dimensional vector representing probabilities of conservation against mutations to 20 different amino acids. Thus, for a sequence of length *N*, a *Nx20* matrix called PSSM is formed. PSI-BLAST search was performed against a filtered non-redundant sequence database, Uniref100 with 3 iterations by using "blastpgp" program.

For the *i*th position residue of a sequence, X_{ia} attribute is calculated by

$$X_{ia} = \frac{1}{w}\sum_{j=w_s}^{w_f} M_{ja} \tag{7}$$

where M_{ja} is the PSSM element at residue (row) j for the column of amino acid, a. The procedure is given in Figure 1.

2.3 Performance Assessment

There are many measures for quantifying prediction accuracy in a binary classification problem, herein order and disorder. Four widely used indices can be given as *Sensitivity, Specificity, Accuracy and Matthews' Correlation Coefficient* (Melamud and Moult, 2003; Su *et al.*, 2006). The definitions of these measures are given by the following equations:

$$\begin{aligned}
&Accuracy\ (Acc) = \frac{(TP+TN)}{(TP+FP+TN+FN)}\\
&Sensitivity\ (Sens) = \frac{TP}{TP+FN}\\
&Specificity\ (Spec) = \frac{TN}{TN+FP}\\
&Matthews'\ Correlation\ Coefficient\ (Mcc) =\\
&\frac{(TP\times TN - FP\times FN)}{\sqrt{(TP+FP)\times(FP+TN)\times(TN+FN)\times(FN+TP)}}
\end{aligned} \tag{8}$$

where *TP* (true positive) is the number of correctly classified disordered residues, *TN* (true negative) is the number of correctly classified residues which have the ordered structure, *FP* (false positive) is the

number of ordered residues incorrectly classified as disordered and *FN* (false negative) is the number of disordered residues incorrectly classified as ordered. Therefore *sensitivity* (Sens) and *specificity* (Spec) represent the fraction of correctly identified residues as disorder and order, respectively (Wu and Mclarty, 2000). The correlation coefficient, introduced by Matthews, gives 1, 0 and -1 for perfect, random and completely wrong predictions, respectively.

Although the *Matthews' correlation coefficient* provides a much more balanced evaluation of prediction than the percentages, all three measures are critically affected by the relative frequency of the target, and they are not suitable for isolated evaluation. For example, in a situation in which all residues are estimated to be disordered, taking a value of *sensitivity* equal to 1 yields a *specificity* equal to 0. Then, accuracy is rather affected by the relative frequencies of the two classes. A *sensitivity* of less than 50% but a *specificity* of more than 80% demonstrates an *under-prediction* of a predictor which has the tendency of predicting order more than disorder.

Therefore, *probability excess* has been recommended as an unbiased measure for evaluating the performance of prediction (Yang *et al.*, 2005). *Probability excess* is independent of the relative class frequencies by means of the evaluation of *sensitivity* and *specificity* values cooperatively with *sensitivity* + *specificity* − 1, that can be graphed by a plot of *sensitivity* versus *specificity*. It is defined by

$$Probability\ Excess\ (ProbEx) = \frac{TPxTN - FPxFN}{(TP + FN)x(TN + FP)} \tag{9}$$

The values of greater than 0.5 reveal an acceptable prediction performance in *probability excess* criteria. Here the value of 1 is also an indicator of a perfect predictor. When the properties of all aforementioned evaluation measurements are considered, it can be said that the *probability excess* can be assumed as a more effective evaluation criterion. Accordingly it was the preferred measure to evaluate the success rate of the methods in our study.

2.4 Computational Networks

A major aim of the study is to find an efficient computational method that can provide information about the structural class among ordered and disordered proteins concerning different biochemical and physical features of amino acids.

For this purpose, a new algorithm is proposed in the study, named Border Vector Detection and Extended Adaptation (BVDEA) to achieve an accurate prediction of disordered data. In addition, three oth-

er methods, Generalized Regression Neural Network (GRNN), Learning Vector Quantization (LVQ), Border Vector Detection and Adaptation (BVDA) were also carried out for comparison.

Border Vector Detection and Extended Adaptation

The algorithm of the Border Vector Detection and Adaptation (BVDA) is based on partitioning of the feature space by using reference vectors selected from the training set (Kasapoğlu and Ersoy, 2007). The process which is named as border vector detection forms the initial step. In the next step, the selected vectors are adapted in a way that is similar to the LVQ (Learning Vector Quantization) method (Bishop, 1995). The proposed method, Border Vector Detection and Extended Adaptation (BVDEA) is based on a similar way of border vector detection offered by the BVDA method but an alternative way for adaptation process of the vectors.

In the first stage of the method, the detection process of border vectors occurs. Firstly, class centers are assigned as the initial border vectors. The samples which are the nearest to the class mean are determined as class centers. The training set $\{(\boldsymbol{x}_1, y_1),\ (\boldsymbol{x}_2, y_2),\ \dots,\ (\boldsymbol{x}_n, y_n)\}$, has n number of samples and m number of class labels, $\boldsymbol{x}_i \in \mathbb{R}^N$, $y_i \in \{1,2,...,m\}$.

Class means $\boldsymbol{o}_i$, are presented as an arithmetic mean of the training samples belonging to the related class. By using the obtained means, class centers are determined as

$$\boldsymbol{c}_i = \boldsymbol{x}_k, \begin{cases} k = \arg\min\{\boldsymbol{D}_j\} \\ \boldsymbol{D}_j(\boldsymbol{o}_i, \boldsymbol{x}_j) = \sum_{d=1}^{N} \sqrt{(o_i(d) - x_j(d))^2},\ \{x_j \,|\, y_j = i\} \end{cases} \tag{10}$$

where $(1 \le i \le m)$, $(1 \le j \le n)$.

The attained class centers constitute the initial border vector set for all classes. Thus, the initial border vector set, $\mathbf{B_0}$ is given by $\mathbf{B_0} = \{(\boldsymbol{c}_1, y_1),\ (\boldsymbol{c}_2, y_2),\ \dots,\ (\boldsymbol{c}_m, y_m)\}$. $\mathrm{m}_0 = m$ denotes the number of the members of the initial border vector set. $\mathbf{B}_0^i$ denotes $(\boldsymbol{c}_i, y_i)$.

The initial border vector set constitutes the initial reference vectors for all classes. In the beginning, the reference vector set of class i is defined as $\mathbf{R}_i(0) = \mathbf{B}_0^i,\ i = 1,2,...,m$. During the detection process, new border vectors are added to these reference vectors. At the end of the process, the reference vector set obtained for class i becomes $\mathbf{R}_i = \mathbf{B}_0^i \cup \mathbf{B}_i$. The number of members of class i reference vector set is de-

termined as $1+\mathrm{m}_i$. $\mathbf{B}_i$ is the added border vector set and m_i is the number of its members.

In the detection procedure, all the training samples are processed iteratively. Any of the input samples are compared with the given reference vectors of its class. If the input training sample $(\boldsymbol{x}_k, y_k)$ is not in the same class with the nearest reference vector $\boldsymbol{b}_w$, then this sample is added to the reference vectors, $\mathbf{R}_{i=q}(\mathrm{t}) = \mathbf{R}_{i=q}(\mathrm{t}-1) \cup \{(\boldsymbol{x}_k, y_k)\}$ and the process is continued with the next sample. The detection process of the nearest reference vector is given as follows:

$$\begin{aligned} &\boldsymbol{D}_j(\boldsymbol{x}_k, \boldsymbol{b}_j) = \left\| \boldsymbol{x}_k - \boldsymbol{b}_j \right\|, \;\; j = 1, ..., (m + \mathrm{m}_\mathrm{t}) \\ &w = \arg\min \{ D_j \} \end{aligned} \tag{11}$$

where m_t represents the number of the border vectors that are added until time t. At time $\mathrm{t}=0$, when the adaptation process starts, the border vector set $\mathbf{B}$ is the combination of the border vectors and the center vectors of all the classes (12). The attained vector set $\mathbf{B}$ is used for the partitioning of the feature space.

$$\mathbf{B} = \bigcup_{0 \le i \le m} \mathbf{B}_i \tag{12}$$

The border vectors are obtained as $\mathbf{B} = \{(\boldsymbol{b}_1, y_1), (\boldsymbol{b}_2, y_2), \ldots, (\boldsymbol{b}_{nb}, y_{nb})\}$ where *nb* is the total number of border vectors.

In the adaptation stage, all samples are compared with existing border vectors, $\mathbf{B}$ and the winner vector, $\boldsymbol{b}_w$ is determined according to the nearest neighbor rule. If the winner vector does not have the same label with the sample, it is pushed away. On the other hand, the nearest vector which is in the class of input sample is brought closer to the input vector by

$$y_j \neq y_{bw} \Rightarrow \begin{cases} \boldsymbol{b}_w(\mathrm{t}+1) = \boldsymbol{b}_w(\mathrm{t}) - \eta(\mathrm{t}).(\boldsymbol{x}_j - \boldsymbol{b}_w(\mathrm{t})) \\ \boldsymbol{b}_{y_j}(\mathrm{t}+1) = \boldsymbol{b}_{y_j}(\mathrm{t}) + \eta(\mathrm{t}).(\boldsymbol{x}_j - \boldsymbol{b}_{y_j}(\mathrm{t})) \end{cases} \tag{13}$$

where learning rate $\eta(\mathrm{t})$ is a decreasing function in time. In the experimental work, the most appropriate value was obtained by attempting different learning rate coefficients. If the winner vector has the same label with the given sample, it is awarded by being brought closer as given by

$$y_j = y_{bw} \Rightarrow \{\boldsymbol{b}_w(\mathrm{t}+1) = \boldsymbol{b}_w(\mathrm{t}) + \eta(\mathrm{t}).(\boldsymbol{x}_j - \boldsymbol{b}_w(\mathrm{t})) \tag{14}$$

In BVDEA, counter to BVDA, at least one border vector is ensured to be rewarded in each repetition. This approach shortens the learning time and provides an improvement of the generalization capability of the algorithm. Generation of additional reference vectors used in BVDA during adaptation was skipped from the method BVDEA because supplementing the number of reference vectors during training may cause over-fitting and increase training time.

3 RESULTS AND DISCUSSION

In the problem of predicting disordered regions, the first step for applying the machine learning techniques is to prepare the dataset. For this purpose, a balanced data set was constituted from completely ordered and completely disordered proteins in this work. Then, the input pattern for each residue of the dataset was obtained by using the sliding windows with length equal to 21. Each pattern contained 120 attributes comprised of 50 composition-based attributes, 50 property-based attributes and 20 evolution-based attributes. Patterns are labeled as 0 or 1 if the class label of the corresponding residue is order or disorder, respectively.

The aim of the study was to test the success of the prediction of the proposed method together with chosen optimum parameters and to compare the results with a few known methods. For this purpose, the proteins of the dataset were divided into 6 different subsets whose sequence lengths were approximately equal. Each subset included the balanced number of order and disorder residues. One of these subsets was allocated as validation set for choosing the optimum parameters. Remaining 5 subsets were used for training the methods via 5 fold cross validation. Therefore, at each time, the different combinations of 4 subsets were used for training while the remaining subset was used for testing. The average of the success rates were obtained from 5 repetitions of training/testing. Thus, any dependency on some initializations such as detection of border vectors and order of training samples on determining the prediction success was minimized by means of the *k*-fold cross validation technique.

3.1 Evaluation Applications

In this work, initially, the obtained dataset COD159 was randomly partitioned into two parts as training and evaluation set to pre-evaluate the optimum learning rate parameter η of the proposed method. Thus, the BVDEA was trained for η parameter values between 0.1 and 0.9 at intervals of 0.1, and also for the smallest values 0.01, 0.001, 0.0001. During training, the prediction results both on training and evaluation sets were recorded at every 50 adaptations. For each candidate parameter, the maximum *probability*

excess (*probEx*) values of the evaluation and training sets were found from among 12000 records. Besides, the *probEx* value of the evaluation set at the time of the maximum success on training self-consistency was assigned as the testing result of training with the corresponding η parameter. The maximum and the testing *probEx* values of the evaluation set are given in Table 2.

According to the testing results of the evaluation set, the best learning rate was estimated as 0.5 or 0.9 for further investigation. However, using a constant rate during training causes oscillation in success values, especially at high rates. Figure 2 shows the oscillating prediction results of both evaluation and training sets for η equal to 0.5. Hence, it has been necessary to use an exponentially descending function of time for learning rate as

$$\eta(t) = \eta_0 e^{-t/\tau} \tag{15}$$

where τ parameter is added to the system. Kasapoğlu et al. determined the optimum values for the pair of η-τ parameters as 0.1-1000. For complex problems, the pair of 0.2-6750 was suggested. Nevertheless, since the appropriate values of η was determined as 0.5 and 0.9 for BVDEA, several combinations of these two values (η-τ) were evaluated in this study. The evaluated values are given as $\eta \in \{0.5, 0.9\}$ and $\tau \in \{6750, 9500, 11500, 13500, 15000\}$.

The evaluations were performed after 5 cross training and testing applications for the reorganized data that was mentioned above.

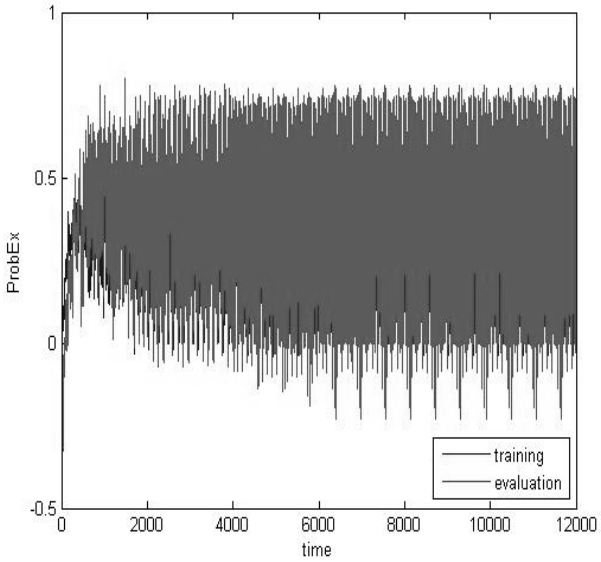


Fig. 2. Prediction results of both evaluation and training sets at $\eta = 0.5$.

Table 2. The maximum and the testing *probability excess* values of the evaluation set.

η	***probEx* (max)**	***probEx* (test)**
0.0001	0.3709	0.3688
0.001	0.6374	0.6281
0.01	0.6612	0.5620
0.1	0.6674	0.5320
0.2	0.7562	0.6529
0.3	0.7831	0.6322
0.4	0.8006	0.7324
0.5	0.8016	0.7366
0.6	0.8071	0.6915
0.7	0.7758	0.6353
0.8	0.7872	0.6405
0.9	0.8192	0.6901

For choosing candidate parameter pairs, the success values of all testing subsets were taken into consideration, concurrently, to make a general evaluation. The maximum success rates in *probability excess* for all testing sets were obtained at each parameter pair. The values are given in the following tables for each subset. It is noted that these maximum values are not accepted as prediction success of the method. The test successes of the method are determined after evaluation and validation applications. By examining the results given in Table 3, the 0.5-9500, 0.5-13500, 0.9-13500 and 0.9-15000 η-τ pairs were determined as candidate parameters for the method.

Any other general evaluation was verified by investigating the training behavior for chosen parameter pairs. For this purpose, for each candidate pair, the prediction results of both testing and training sets of each subset for the chosen η-τ parameters, 0.5-9500, 0.5-13500, 0.9-13500 and 0.9-15000 were graphed regarding the *probEx* values of 2000 records (Appendix B). Each figure contains 5 plots for 5 subsets. The plots of training and testing for the pairs 0.5-9500 are also given here as an example (Figure 3).

By examining the graphs presented in Figure 3, it is observed that the curves of the testing success for the η-τ pair 0.5-9500 changes their direction at 450-500 levels of the records i.e. while the training success goes on increasing, the testing success begins to decrease at these levels, and the two curves start to separate from each other. Correspondingly, it was decided that the training for this η-τ pair had to go on

up to level ~500 at most. For the other rate pairs, the corresponding graphs were also examined, and this limitation was determined as ~550 for 0.5-13500 and ~700 for 0.9-13500, heuristically. On the other hand, in the graphs for the parameter pair of 0.9-15000, the regions of separation for the curves exhibit significant diversity. While the level is approximately 500 in the curve of subset 4, it is almost 1000 for subset 5 (see Appendix B for more details). This precludes performing a generalization, so the pair 0.9-15000 was eliminated from the evaluations.

After the general evaluations, validation applications were performed to determine the parameter pairs belonging to the related subset. In order to make a decision, the training of a subset with all candidate parameter pairs was continued until predefined times (record level) and then, up to this time, we investigated at which time the training set has the maximum success value i.e. the maximum self-consistency result of training in *probEx*.

The results attained from testing of validation set were obtained according to the decision regions described by the border vectors at stopping points of trainings. The parameter pair that gave the most successful validation result for the corresponding subset was then assigned to the associated subset.

Table 3. The maximum *probability excess* values of testing subsets at given η-τ pairs.

η-τ	***probEx (max)***				
	Subset 1	**Subset 2**	**Subset 3**	**Subset 4**	**Subset 5**
0.5-6750	0.6426	0.4317	0.5685	0.4801	0.6320
0.5-9500	0.6767	0.4338	0.6148	0.5295	0.6237
0.5-11500	0.6798	0.3971	0.6097	0.4404	0.6062
0.5-13500	0.6746	0.4884	0.5984	0.4415	0.6072
0.5-15000	0.6808	0.4160	0.5911	0.4415	0.6340
0.9-6750	0.6333	0.3309	0.6107	0.4619	0.5876
0.9-9500	0.6467	0.4275	0.6004	0.4662	0.6144
0.9-11500	0.6653	0.4191	0.5788	0.5274	0.6155
0.9-13500	0.7180	0.4160	0.5644	0.4984	0.6144
0.9-15000	0.6777	0.3981	0.5994	0.5134	0.6515

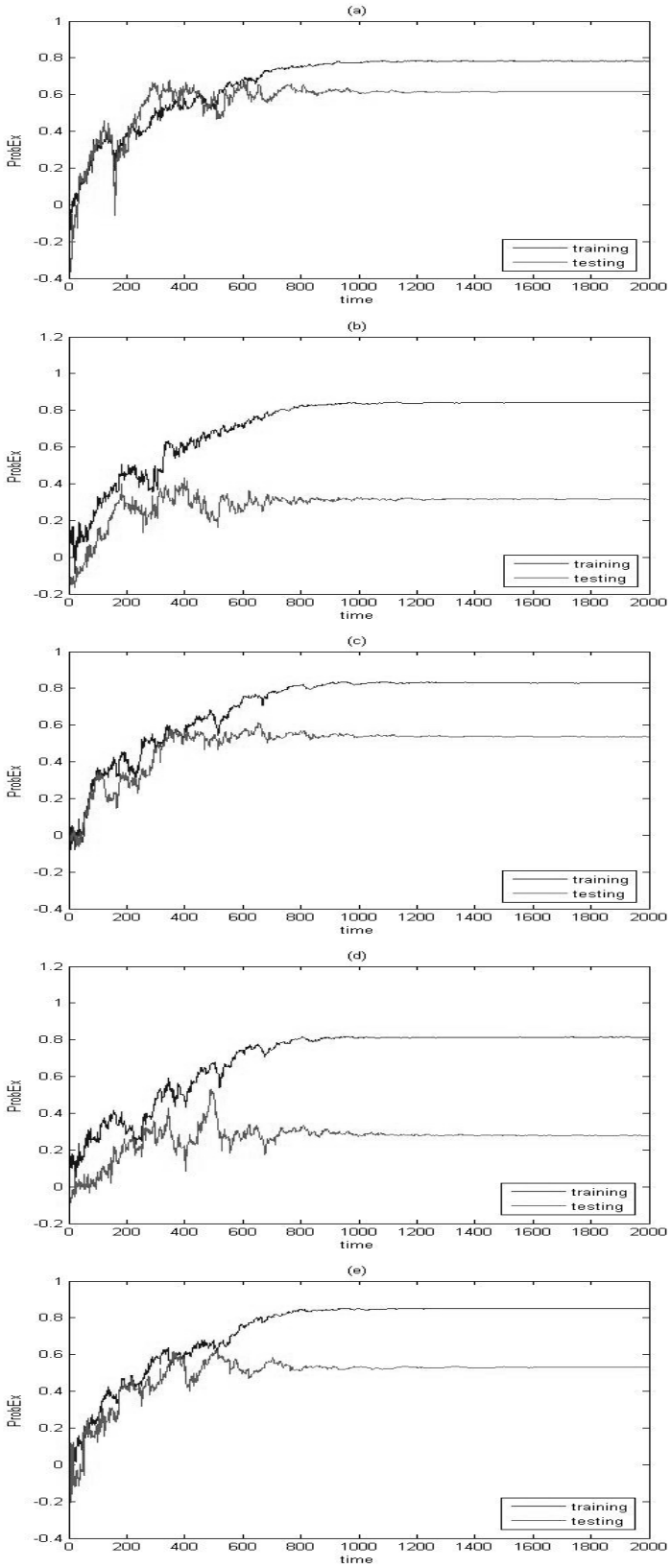


Fig. 3. Prediction results of both testing and training sets with $\eta - \tau = 0.5 - 9500$ (a) for subset 1 (b) for subset 2 (c) for subset 3 (d) for subset 4 (e) for subset 5.

The validation results of training BVDEA with 3 candidate parameter pairs, obtained for each subset, are given in Appendix C. The optimum parameter pairs selected for subsets and the stopping times of training with related parameter pairs determined by the self-consistency results are given for each subset in Table 4. The stopping times are given in terms of record number.

The classification of testing subsets with BVDEA was accomplished by the nearest distance rule to the border vectors obtained during training.

3.2 Evaluation Applications

The performance of BVDEA on predicting disordered regions of a protein was compared with several methods. LVQ and BVDA were preferred due to their similar learning approach with BVDEA. Besides, the method of GRNN was included.

Before determining the success values for the chosen methods, some evaluation and/or validation applications were executed for choosing associated parameters.

A series of runs as in BVDEA were generated for the method BVDA. At first, in all subsets, trainings were realized for different η-τ pairs. On this occasion, the candidate parameter pairs defined in BVDEA was compared with suggested values by Kasapoğlu and Ersoy (Kasapoğlu and Ersoy, 2007). The evaluated values were given as η-$\tau \in$ {0.1-1000, 0.2-6750, 0.5-9500, 0.5-13500, 0.9-13500}. The maximum success rates in terms of probability excess for testing the subsets at each parameter pair are given in Appendix D. It is observed that the optimum parameter values given in BVDEA cause more successful results than the suggested values in Kasapoğlu and Ersoy's article. Afterwards, the prediction results of both testing and training sets of subset 1, subset 3 and subset 5 with associated candidate η-τ parameters, 0.9-13500, 0.5-13500 and 0.5-9500, respectively, were graphed in terms of the probability excess values of 2000 records as examples for evaluation (see Appendix E). It was concluded that the limiting levels for the training parameter pairs were similar to the levels found in BVDEA.

However, some differences stand out in results. For example, in BVDA, adding new vectors during training causes a continual increase in training success rate, and the success rate approaches 1. Furthermore, testing success usually gets its maximum value at earlier times of training. But, due to the oscillations, the success rates fluctuate between very high and very low values at around these regions. This makes it impossible to stop the training around these regions. As a result, because of the similarity in general evaluation results for both methods, the optimum parameters found in BVDEA for the subsets

were preferred. Anyway, using any other parameter values did not change the results significantly due to the closeness of the success rates with different parameter values (see Appendix D)

For attaining prediction results of BVDA, prediction was performed according to the training settings with the optimum η-τ pairs given in Table 4, and the stopping times of trainings with related parameter pairs determined by the self-consistency results for the subsets as 694, 545, 469, 499 and 498, respectively.

In the LVQ method, the number of codebook vectors (n_c) was considered as a parameter. In this study, two values of n_c were validated. One of them was determined as 50 because BVDEA was performed with approximately 50 vectors; but for the other, the smaller value of 10 was used. The training in LVQ was carried on until 100 epochs.

It was observed that the most successful learning rates were achieved with n_c equal to 10 for all subsets (see appendix F). Thus, 10 codebook vectors were used for training with LVQ. The prediction results of LVQ for each subset were obtained by using the optimum number of vectors for subset trainings.

The learning rate parameter equal to 0.01 was used during learning. It was demonstrated that the prediction results were badly affected for greater values of this parameter. Testing results with validation set with the learning rate parameter equal to 0.05 and with 10 vectors are given in Appendix F as an example.

Similarly, the GRNN was also trained with two different sigma (σ) parameter values. Validation applications were verified with $\sigma \in \{0.3, 0.7\}$. The value, 0.3 was suggested in previous studies (Ersöz *et al*., 2004). The values obtained for all subsets are listed in Appendix G. σ equal to 0.7 was found to give maximum success for all subsets. The classification of testing subsets was achieved by using the optimum σ value of 0.7 determined during training.

The success rates with the testing sets of the methods with the given settings are presented in Table 5. The performances are provided in terms of *probability excess*.

Table 4. The optimum η-τ pairs and the stopping times of BVDEA for each subset.

Subset 1	**Subset 2**	**Subset 3**	**Subset 4**	**Subset 5**
0.9-13500	0.5-13500	0.5-13500	0.5-9500	0.5-9500
604	503	498	494	494

Table 5. The testing *probability excess* values for four methods.

Methods	**Subset 1**	**Subset 2**	**Subset 3**	**Subset 4**	**Subset 5**
BVDEA	0.7128	0.4811	0.5901	0.5177	0.5948
BVDA	0.6198	0.3372	0.5427	0.3974	0.5103
LVQ	0.7820	0.4412	0.5417	0.5627	0.4969
GRNN	0.7180	0.4034	0.6220	0.5134	0.5474

Table 6. Comparison of the performances of four prediction methods.

Methods	***Sens***	***Spec***	***Acc***	***Mcc***	***ProbEx***
BVDEA	0.7964	0.7850	0.7907	0.5858	0.5814
BVDA	0.6981	0.8707	0.7844	0.5818	0.5688
LVQ	0.7263	0.8345	0.7804	0.5714	0.5608
GRNN	0.7309	0.7506	0.7408	0.4878	0.4815

Consequently, the overall performances of all methods were calculated by averaging the results of 5 subsets. The success values in terms of all measures for all methods are listed in Table 6. The given list is ordered with respect to the *probability excess* measures.

The results with the testing data indicate that BVDEA achieves the best performance in prediction with the *ProbEx* of 0.5814. The nearest performance belongs to the method of LVQ. However it provides an unbalanced prediction between order and disorder residues with high *specificity* but relatively low *sensitivity*. Overall, the results support BVDEA as a successful classifier for predicting disorder and order.

3.3 Comparison with Existing Tools

For the second comparison, the performance of specific disorder prediction tools, PONDRs, DisEMBL, GlobPlot, DISOPRED2, FoldIndex, RONN, DisPRO, PreLink and DisPSSMP were investigated. These methods have been presented in the literature as publicly available web servers or packages. The server for DisEMBL provides three choices for prediction as mentioned before. The three applications are given as DisEMBL (hot), DisEMBL (465) and DisEMBL (coils). The results for the methods were collected from the work of Yang et al. and the work of Su et al (Yang *et al*., 2005; Su *et al*., 2006)

The performance of BVDEA on COD159 dataset was compared with the 11 methods in Table 7. Similarly, the performance of the methods on blind testing set, D80, were also examined (Table 8). In Table

7 and Table 8, the results were compiled with all measures, *sensitivity*, *specificity*, *accuracy*, *Matthews' correlation coefficient* and *probability excess*. The success order of the algorithms was given with regard to the favored measure, *probability excess*.

The prediction performances on testing the balanced set of COD159 were also compared with the given methods via the graph of probability excess given in Figure 4. Inspection of the results exhibits that while DisPSSMP performs slightly better than BVDEA, BVDEA gives slightly more balanced prediction compared with DisPSSMP.

The four methods, BVDEA, DisPSSMP, RONN and FoldIndex significantly performed better than the other methods. Most of them are found in the left side of the trapezoid graph of Figure 4. This shows the tendency of predicting order i.e. under-prediction of disorder.

For obtaining prediction results of BVDEA for blind testing set of D80, the classification was performed by testing the set with the border vectors obtained during training (Table 4). The testing performances for each training application were averaged, and this gives the overall success results for D80 blind set (Table 8). All performances are presented in Figure 5.

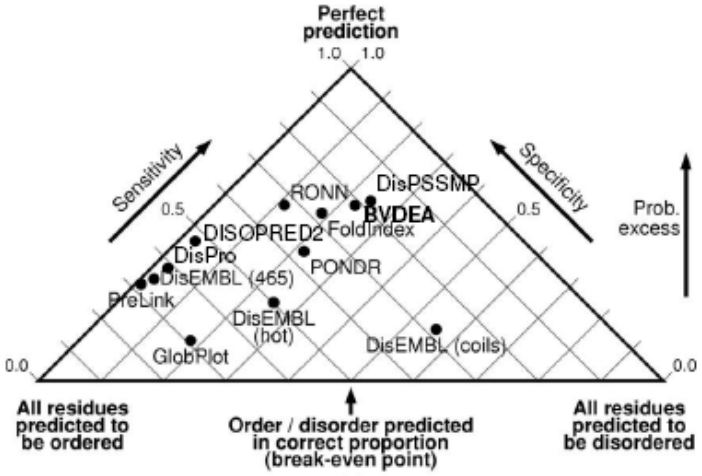
Table 7. Comparing the performance of BVDEA on testing data COD159 with eleven existing tools.

Methods	*Sens*	*Spec*	*Acc*	*Mcc*	*ProbEx*
DisPSSMP	0.825	0.765	0.795	0.589	0.590
BVDEA	0.796	0.785	0.791	0.586	0.581
RONN	0.675	0.888	0.782	0.580	0.563
FoldIndex	0.722	0.815	0.769	0.540	0.536
DISOPRED2	0.469	0.981	0.725	0.543	0.449
PONDR	0.632	0.782	0.707	0.420	0.414
DisPro	0.383	0.982	0.683	0.467	0.365
DisEMBL(465)	0.348	0.978	0.663	0.430	0.327
PreLink	0.319	0.991	0.655	0.430	0.310
DisEMBL(hot)	0.502	0.749	0.626	0.260	0.251
DisEMBL(coil)	0.719	0.446	0.583	0.170	0.165
GlobPlot	0.308	0.821	0.565	0.151	0.129

Table 8. Comparing the performance of BVDEA on blind testing data D80 with eleven existing tools.

Methods	*Sens*	*Spec*	*Acc*	*Mcc*	*ProbEx*
DisPSSMP	0.767	0.848	0.808	0.463	0.615
BVDEA	0.817	0.728	0.773	0.451	0.545
RONN	0.603	0.878	0.741	0.395	0.481
DisPro	0.418	0.993	0.706	0.578	0.411
DISOPRED2	0.405	0.972	0.689	0.470	0.377
PONDR	0.557	0.816	0.687	0.278	0.373
DisEMBL(hot)	0.492	0.840	0.666	0.260	0.332
DisEMBL(465)	0.334	0.981	0.658	0.437	0.315
FoldIndex	0.488	0.811	0.650	0.224	0.299
PreLink	0.237	0.947	0.592	0.219	0.183
GlobPlot	0.372	0.811	0.592	0.140	0.183
DisEMBL(coil)	0.740	0.424	0.582	0.104	0.165

Fig. 4. The performances of twelve predictors on testing data, COD159.



As seen from the results, the BVDEA gives rather high accuracy even with the dominance of ordered residues in the blind test set, with closest results to the best performing DisPSSMP among the others. Both are the only methods that have a probability excess value over 0.5. On the other hand, the BVDEA perform best in predicting disorder with the highest value of specificity at 82%. The DisEMBL(coils) has a good performance of sensitivity but reveals a significant over-prediction.

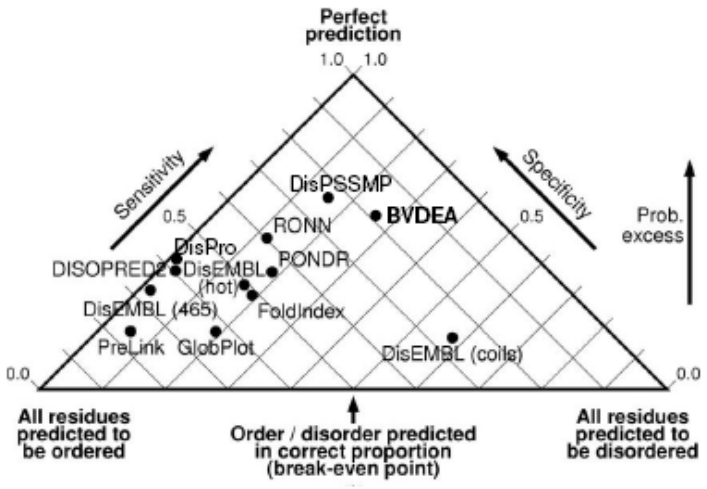


Fig. 5. The performances of twelve predictors on blind testing data, D80.

For the other methods, they yield very high scores of specificity but they attain this at the expense of missing estimate with a significant number of disordered residues. This means that under-prediction occurs.

4 CONCLUSIONS

Studies on structural genomics indicate that numerous protein segments remain unfolded in their native states. Contrary to the structure – function paradigm, these regions fail to fold into a fixed 3D structure under physiological condition yet exhibit function. Currently, these proteins are generally referred to as "natively unfolded" or "intrinsically disordered". Upon noticing that many of these proteins play key role in vital functions and also in some diseases, identification of the disordered regions has became a demanding process for structure prediction and functional characterization of proteins. Therefore, many studies have been motivated on accurate prediction of disorder.

In this study, a novel method was developed for predicting disorder as an alternative accurate classifier. For attesting the performance of the method, three computational learning techniques and eleven specific tools were used for comparison. In order to perform the predictions, first the collected data was prepared by utilizing a variety of structural features for proteins. Then, training was executed based on the data by 5-fold cross validation.

When compared with the three learning methods of GRNN, LVQ and BVDA, the proposed method BDVEA gives the best accuracy on classification. Here, BDVEA provides more fast and robust learning

as compared to the others. The training time is almost the same with LVQ (~3 minutes), the half of BVDA and the quarter of GRNN.

Furthermore, comparison with the other existing disorder predictors demonstrates that the method achieves quite good prediction performance in finding disordered regions of proteins. Except for DisPSSMP, BVDEA outperforms all other ten methods significantly, without either under-predicting or over-predicting the disordered regions. This is confirmed with both main and blind testing successes. Nevertheless, for testing the main set, BVDEA yields roughly equal performance with DisPSSMP and more balanced prediction accuracy. Besides, it reveals a highest score of sensitivity on prediction of blind test set as compared to all other methods, including DisPSSMP. As evident from the results, BVDEA can be suggested as a competitive method to achieve accurate predictions of disordered regions. This also provides insights for projects on drug discovery and gene finding.

Funding: This research was supported by TUBITAK-TBAG 104T505 and Cukurova University Research Foundation MMF2006D12.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI - BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., 25, 3389-3402.

Bishop,C.M. (1995) *Neural Networks for Pattern Recognition*. Oxford University Press Inc., New York, 482.

Cheng,J. *et al*. (2005) Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. *Data Min. Knowl. Dis*., 11, 213–222.

Coeytaux,K. and Poupon, A. (2005) Prediction of Unfolded Segments in a Protein Sequence Based On Amino Acid Composition. *Bioinformatics,* 21(9), 1891–1900.

Dosztányi,Z. *et al*. (2005) The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates Between Folded and Intrinsically Unstructured Proteins. *J. Mol. Biol.*, 347, 827–839, 2005.

Dunker,A.K. *et al*. (1997) On the Importance of Being Disordered. *PDB Newsletter*, 81, 3-5.

Dunker,A.K. *et al*. (2000) Intrinsic Protein Disorder in Complete Genomes. *Genome Informatics*, 11, 161-171.

Dunker,A.K. *et al*. (2001) Intrinsically Disordered Protein. *Journal of Molecular Graphics and Modeling*, 19, 26-59.

Ersöz,İ. *et al*. (2004) Secondary Structure Prediction of Hemoglobin by Neural Networks. *ANNIE 2004*, St. Louis, Missouri, USA, 2004.

Fischer,E. (1894) Einfluss Der Configuration Auf Die Wirkung Der Enzyme. *Ber. Dt. Chem. Ges.*, 27, 2985–2993.

Galzitskaya,O.V. *et al*. (2006) Prediction of Amyloidogenic and Disordered Regions in Protein Chains. *PLoS Bioinf*., 2 (12), 1639-1648.

Garbuzynskiy,S.O. *et al*. (2004) To Be Folded or To Be Unfolded? *Protein Sci*., 13, 2871-2877.

Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, 3, 522-524.

Jones,D.T. and Ward,J.J. (2003) Prediction of Disordered Regions in Proteins from Position Specific Scoring Matrices. *Proteins*, 53, 573-578.

Kasapoglu,N.G. and Ersoy,O.K. (2007) Border vector detection and adaptation for classification of multispectral and hyperspectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 45, 12, 3880-3893.

Koshland,D.E. (1958) Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci*., USA, 44, 98–104.

Li,X. *et al*. (1999) Predicting Protein Disorder for N-, C- and Internal Regions. *Genome Inform Ser. Workshop Genome Infor*., 10, 30-40.

Li,X. *et al*. (2000) Comparing Predictors of Disordered Protein. *Genome Informatics*, 11, 172-184.

Linding,R. *et al*. (2003a) Protein Disorder Prediction: Implications for Structural Proteomics. *Structure*, 11(11), 1316-1317.

Linding,R. *et al*. (2003b) Globplot: Exploring Protein Sequences for Globularity and Disorder. *Nucleic Acids Research*, 31(13), 3701–3708.

Melamud,E. and Moult,J. (2003) Evaluation of Disorder Predictions in CASP5. *Proteins*, 53, 561-565.

Mirsky,A.E. and Pauling,L. (1936) On the Structure of Native, Denatured, and Coagula ted Proteins. *Proc. Natl. Acad. Sci*., USA, 22(7), 439–447.

Obradovic,Z. *et al*. (2003) Predicting Intrinsic Disorder from Amino Acid Sequence. *Proteins: Structure, Function, and Genetics*, 53, 566-572.

Peng,K. *et al*. (2005) Optimizing Long Intrinsic Disorder Predictors with Protein Evolutionary Information. *Journal of Bioinformatics and Computational Biology*, 3(1), 35-60.

Peng,K. *et al*. (2006) Length-Dependent Prediction of Protein Intrinsic Disorder. *BMC Bioinformatics*, 7 (1), 208.

Prilusky,J. *et al*. (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics,* 21(16), 3435-3438.

Qian,N. and Sejnowski,T.J. *et al*. (1988) Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *J. Mol. Biol*., 202, 865-884.

Radivojac,P. *et al*. (2004) Protein Flexibility and Intrinsic Disorder. *Protein Science*, 13(1), 71-80.

Romero,P. *et al*. (1997) Identifying Disordered Regions in Proteins from Amino Acid Sequence. *In Proc. IEEE Int. Conf. On Neural Networks*, 90-95.

Romero,P. *et al*. (2001) Sequence Complexity of Disordered Protein. *Proteins: Structure, Function and Genetics,* 42, 38-48..

Shimizu,K. *et al*. (2004) Predicting The Protein Disordered Region Using Modified Position Specific Scoring Matrix. *The 15th International Conference on Genome Informatics,* 150.

Shimizu,K.Y. *et al*. (2007) Predicting Mostly Disordered Proteins by Using Structure-Unknown Protein Data. *BMC Bioinformatics*, 8, 78.

Su,C. *et al*. (2006) Protein Disorder Prediction by Condensed PSSM Considering Propensity for Order or Disorder. *BMC Bioinformatics*, 7, 319.

Tompa,P. *et al*. (2002) Intrinsically Unstructured Proteins. *Trends in Biochemical Sciences*, 27(10), 527-533.

Uversky,V.N. *et al*. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, 41, 415-427.

Vucetic,S. *et al*. (2003) Flavors of Protein Disorder. *Proteins: Structure, Function and Genetics*, 52, 573-584.

Ward,J.J. *et al*. (2004a) Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol*., 337, 635–645.

Ward,J.J. *et al*. (2004b) The DISOPRED Server for the Prediction of Protein Disorder", Bioinformatics, 20, 2138–2139.

Weathers,E.A. *et al*. (2004) Reduced Amino Acid Alphabet is Sufficient to Accurately Recognize Intrinsically Disordered Protein. *FEBS Letters*, 576, 348–352.

Williams,R.M. *et al*. (2001) The Protein Non-Folding Problem: Amino Acid Determinants of Intrinsic Order and Disorder. *Pacific Symposium on Biocomputing*, 89-100.

Wootton,J.C. and Federhen,S. (1993) Statistic of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, 17, 149-163.

Wright,P.E. and Dyson,H.J. (1999) Intrinsically Unstructured Proteins: Re-assessing the Protein Structure-Function Paradigm. *J. Mol. Biol*., 293, 321-331.

Wu,C.H. and Mclarty,J.W. (2000) *Neural Networks and Genome Informatics*. Elseiver, USA, 8-11, 69-76, 97-99, 116-119.

Yang,R.Z. *et al*. (2005) RONN: The Bio-basis Function Neural Network Technique Applied to the Detection of Natively Disordered Regions in Proteins. *Bioinformatics*, 21, 3369-3376.